Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5894–5905
Marseille, 20-25 June 2022

An automatic model and Gold Standard for translation alignment of Ancient Greek

Tariq Yousef*, Chiara Palladino†, Farnoosh Shamsian*
Anise d'Orange Ferreira⋆, Michel Ferreira dos Reis⋆
*University of Leipzig
Augustusplatz 10, 04109 Leipzig, Germany
tariq.yosef@uni-leipzig.de, farnoosh.shamsian@uni-leipzig.de

†Furman University
3300 Poinsett Highway, 29613, Greenville SC, USA
chiara.palladino@furman.edu

⋆Universidade Estadual Paulista (UNESP)
Rod. Araraquara-Jaú Km 1 - Bairro dos Machados Machados - Araraquara/SP - CEP 14800-901, Brazil
anise.ferreira@unesp.br, michelfereis@yahoo.com.br

Abstract

This paper illustrates a workflow for developing and evaluating automatic translation alignment models for Ancient Greek. We designed an annotation Style Guide and a gold standard for the alignment of Ancient Greek-English and Ancient Greek-Portuguese, measured inter-annotator agreement and used the resulting dataset to evaluate the performance of various translation alignment models. We proposed a fine-tuning strategy that employs unsupervised training with mono- and bilingual texts and supervised training using manually aligned sentences. The results indicate that the fine-tuned model based on XLM-Roberta is superior in performance, and it achieved good results on language pairs that were not part of the training data.

Keywords: Translation Alignment, Gold Standard, Alignment Guidelines, Ancient Greek

1. Introduction

Word alignment is defined as the operation of comparing two or more texts in order to find correspondences between their textual units. When the texts being compared are in different languages (also called parallel texts or parallel corpora), the task is more specifically called translation alignment. The result often takes the form of a list of pairs of items, which can be larger text chunks like documents or paragraphs, but more frequently sentences and words (Kay and Röscheisen, 1993; Véronis, 2000). Translation alignment is a very important task in Natural Language Processing. Since the 90s, many models have been developed to automatically establish correspondences between corpora in different languages (Brown et al., 1993), and parallel corpora are used for a variety of purposes, including neural and statistical machine translation (DeNero and Klein, 2007), automatic bilingual lexicon extraction (Yousef et al., 2022), corpus linguistics (Baker, 2000), language learning (Palladino et al., 2021), and cross-lingual annotation projection (David et al., 2001; Padó and Lapata, 2009; Müller, 2017; Nicolai and Yarowsky, 2019). However, developing reliable models for automatic alignment is a challenging task: cultural, contextual, and linguistic differences make much of the difficulty in establishing perfect correspondences across languages. Therefore, corpora of reliably manually aligned texts or gold standards (and guidelines to create them) are a much-desired resource, since they are often essential to train and evaluate automatic alignment models (Dagan et al., 1999; Graça et al., 2008; Lambert et al., 2005; Mareček, 2008).

To our best knowledge, most current studies on automatic translation alignment and alignment gold standards are conducted on modern languages. In this paper, we will illustrate a workflow for evaluating translation alignment models starting from Ancient Greek texts and translations into English and Brazilian Portuguese. This contribution is structured as follows: we provide a review of the related work on alignment guidelines and gold standards and approaches to automatic word alignment. Next, we describe our work, focusing on creating alignment guidelines for Ancient Greek-English and Ancient Greek-Portuguese developed by domain experts and used to create reliable and high-quality word-level gold standard datasets. The gold standards were evaluated for inter-annotator agreement to ensure reliability. Next, we employ a state-of-the-art alignment workflow that utilizes multilingual contextualized language models and propose a training strategy. The model[1] significantly outperforms the popular statistical models such as Giza++, elforml, and fast_align on Ancient Greek-English; and achieves impressive results even with the absence of training data. In the closing part of the paper, we evaluate the obtained results and propose some lines of future work.

[1] https://github.com/UgaritAlignment/Automatic-Translation-Alignment-of-Ancient-Greek-Texts

2. Related Work

The main purpose of gold standards is to evaluate the performance of automatic alignment models. Creating gold standards requires alignment guidelines that ensure high agreement among annotators, resulting in reliable and accurate results. Therefore, most research on generating gold standards is associated with developing annotation style guides.

There are countless examples of alignment gold standards for pairs of modern European languages, most of which include English: French-English (Melamed, 1998; Och and Ney, 2000), Dutch-English (Macken, 2010), English-Swedish (Holmqvist and Ahrenberg, 2011), Romanian-English (Mihalcea and Pedersen, 2003), Czech-English (Kruijff-Korbayová et al., 2006; Mareček, 2008), English-Spanish (Lambert et al., 2005), and English-Icelandic (Steingrímsson et al., 2021). More rarely, gold standards are produced for several language pairs, e.g. (Graça et al., 2008) for Portuguese, English, French and Spanish. In most cases, a gold standard was created by developing new annotation guidelines or expanding existing ones. Alignments are usually performed by two different annotators on the same corpus, and their agreement is measured to assess the quality of the guidelines, and the reliability of the gold standard. Inter-annotator agreement of the datasets mentioned above ranges from 86.7% for English-Spanish to 96.5% for Spanish-French (Graça et al., 2008).

The sizes of the gold standards used in the literature described above range from 100 sentences per languages pair (Graça et al., 2008), to 1500 sentences (Macken, 2010). Europarl was the main source of parallel data for creating alignment gold standard (Graça et al., 2008; Macken, 2010; Lambert et al., 2005; Holmqvist and Ahrenberg, 2011). Macken (2010), however, used journalistic texts, newsletters, and medical reports to create gold standards for Dutch-English. Most experiments employ the Sure/Possible or Sure/Possible/Null annotation schema (Holmqvist and Ahrenberg, 2011; Kruijff-Korbayová et al., 2006; Graça et al., 2008), while Macken (2010) employed multi-level annotation schema with Regular/Fuzzy/Null as main classes.

Automatic translation alignment is based either on statistical or neural models. The first statistical lexical models for automatic word alignment, known as IBM models, were introduced by Brown et al. (1993). Och and Ney (2003) developed $Giza++$, which has been considered the state-of-the-art in the field for a long time. Later, Dyer et al. (2013) introduced a fast and effective log-linear implementation of IBM Model2, called $fast\_align$, outperforming IBM Model 4. Then, Östling and Tiedemann (2016) proposed $Eflomal$, an efficient and accurate word alignment model using a Bayesian model with Markov Chain Monte Carlo (MCMC) inference. However, all these methods perform poorly in the absence of parallel sentences as training data.

Recently, statistical alignment has been replaced by neural models. Research revealed the possibility of exploiting multilingual contextualized language models to create accurate alignments even without training data. (Dou and Neubig, 2021) introduced *AWESOME* aligner that predicts alignments from similarity matrices, and proposed training objectives to fine-tune language models for better performance. (Jalili Sabet et al., 2020) developed *SimAlign*, an aligner that can produce high-quality word alignments using static and contextualized embeddings. (Stengel-Eskin et al., 2019) introduced a discriminative neural alignment model that leverages similarity matrices of encoder-decoder representations to predict word alignments.

In this paper, we use similarity matrices based on embeddings derived from multilingual contextualized language models such as mBERT and XLM-R and employ multiple alignment extraction approaches introduced by Jalili-Sabet et al (2020) and Dou and Neubig (2021) to predict alignment for Ancient Greek-English parallel texts. Further, we perform supervised and unsupervised fine-tuning of the models using monolingual and bilingual datasets we collected from different resources and high-quality manual alignments available on UGARIT[2].

3. Creating a Gold Standard for Ancient Greek

The Gold Standard created for this research consisted of two datasets of texts manually at word level, in Ancient Greek-English and Ancient Greek-Brazilian Portuguese, with Guidelines developed for this purpose. The corpus was aligned using UGARIT, a platform designed for crowdsourcing efforts in translation alignment of historical and low-resourced languages (Yousef et al., 2022; Palladino et al., 2021; Yousef and Foradi, forthcoming 2022). The materials here described, including Guidelines and datasets, are available at `https://github.com/UgaritAlignment/Alignment-Gold-Standards`.

3.1. The Guidelines

Both the Gold Standard and the Guidelines for Ancient Greek-English were created by two experts who had previously worked with UGARIT. The first draft of the Guidelines was created through multiple meetings, prior to aligning the corpus: the basic structure was developed starting from an already existing model, and considerably expanded to include various language-specific issues that the experts had encountered in the past. Then, the experts aligned a subset of the corpus to test the general consistency and feasibility of the Guidelines: during this phase, for each new issue there was a brief discussion and a preferred annotation style or an improvement in the guide was agreed upon. Each change was incorporated in the final version, and the

[2] http://ugarit.ialigner.com/

alignments were revised accordingly. After the subset was completed, the experts completed the alignment without further discussions, to ensure that the consistency and efficacy of the guidelines could be appropriately tested. The Guidelines for Portuguese were created in a similar fashion: they designed by two domain experts, who also manually aligned the corpus, starting from a previous draft that was expanded substantially by taking the English Guidelines as a model.

The Guidelines consider the types of links allowed by UGARIT, which are one-to-one (1-1), many-to-many (N-N), one-to-many (1-N) and many-to-one (N-1). Links in UGARIT do not include lack of alignment (0 link): words that do not correspond are simply left unaligned. Even though UGARIT does not distinguish between possible and certain alignments, this was addressed differently in the evaluation of inter-annotator agreement (see below). The Guidelines were created specifically with the goal of creating a consistent and reliable Gold Standard to use in machine-actionable models. For this reason, the main structural problem to address was the highly inflected nature of Ancient Greek as a language, which created contrast with the translation languages. We adopted the definition of (Lambert et al., 2005): "the only valid elements in an alignment are single words and indivisible groups of words" (p.275): groups of words are linked together when the meaning of the group is distinct from that of the sequence of each word's meaning, and single words cannot be separated from the rest of the group without changing their meaning (= indivisible lexical unit). Further, we established that correspondence between lexical units had to involve as few words as possible, but as many words as necessary, with the requirement of equivalent meaning between original and translation. Basically, this meant that linguistic structures that were peculiar to Ancient Greek could be aligned as lexical units when necessary, but the general principle allowed an overall prevalence of one-to-one links, which are more useful to train computational models.

3.2. The Corpus

The corpus aligned to develop the Gold Standard included three Ancient Greek texts and translations into English and Portuguese: we included passages from the *Iliad* (2010 words), Plato's *Crito* (1829 words) and Xenophon's *Cyropedia* (1520 words).

The texts chosen provided sufficient diversity of language (Homeric to Koine Greek) and of text genre (poetry, prose, and dialogue), and could be easily aligned at the level of their specific citation unit (sentence, short paragraph, or speaker). In the case of the *Iliad*, we used a text already aligned at sentence level with the English translation currently used in the Perseus dependency treebanks (Bamman et al., 2010). The translations we selected were mostly modern for both English and Portuguese (Murray, 1924; Burnet, 1903; Marchant, 1910; Werner, 2018).

3.3. Further Considerations

The Guidelines revealed aspects that were consistent with similar annotation guides for modern languages. For example, most guidelines address punctuation, omission, phrasal construction and repetition in the same way ours do (punctuation tends not to be aligned; repeated words in only one language are only aligned in the first instance; omission and ellipsis may result in lack of alignment; phrasal constructions, including idioms and proverbial expressions, are considered indivisible units and aligned N-N, and so on and so forth). However, the peculiarities of working with an ancient language often steered our approach in different directions, and required us to think about situations that are simply not that frequent in modern languages. Not only we had to deal with situations that are not often found in languages generally used in alignment (e.g. high inflection of verbs or nouns), but we also had to consider the very inconsistent ways in which certain linguistic and rhetorical structures are addressed, not only across translations, but even within the same translation. So, while the guidelines developed for Spanish and English by (Lambert et al., 2005), which provided the main guiding principles for our own, only included detailed guidance for 7 classes of phenomena, our guidelines include double that (14), and go in much deeper detail to include variations to those: for instance, just our section on determiners includes five separate situations that had to be addressed in detail for English, and six for Portuguese. The status of Ancient Greek as a dead language, where there is only a finite number of texts, also has implications, since it is not possible to verify the accuracy of some things, which requires to make judgement calls in the establishment of translation pairs. For example: can a string be classified as an idiom, if it only appears in one author or one work, or even only once in the entire language corpus? How can we create consistent guidelines for structures, like the genitive absolute, whose semantic function is exclusively established in relation to the context where they appear?

The resulting considerations indicate that it is extremely important to understand which languages are being aligned, and to which cultural tradition they belong. For historical languages, it is simply not possible to obtain well-established, easy-to-align corpora of texts with faithful translations. As a result, guidelines will be necessarily more detailed and, on the other hand, there may be higher chance of disagreement between annotators. However, our research shows that it is possible to create reliable gold standards, provided that strong scholarly expertise in the original language is put to use.

3.4. Alignment Results

Our alignment guidelines do not distinguish between sure and possible alignments as proposed by (Och and Ney, 2003), but when combining the alignments of the two annotators, we defined sure and possible alignment sets for every sentence as follows:

$$S = A_1 \cap A_2 \quad , \quad P = A_1 \cup A_2$$

A_1 and A_2 are the alignments sets created by the first and second annotators, respectively. S denotes sure alignments which include all translation pairs where both annotators agree. P denotes possible alignments where the translation pairs are aligned by at least one annotator.

We exported the gold standards in NAACL Format (Mihalcea and Pedersen, 2003). Each line in the file represents a sentence, a sequence of translation pairs. Each translation pair contains the index of the source token with the index of the target token (zero-indexed) with a hyphen between them followed by a letter (S or P) to indicate if the current alignment is sure or possible alignment.

Inter-Annotator Agreement

Inter-Annotator Agreement (IAA) is a great indicator of the reliability of the annotation guidelines and the quality of the alignment gold standards.

	Grc-Eng	Grc-Por
Sentences	275	183
Grc Tokens	5.359	3.216
Grc Types	2.347	1.587
Eng/Por Tokens	7.515	3.710
Eng/Por Types	1.634	1.355
Sure Alignments	6.240	3.028
Possible Alignments	1.423	864
IAA	86.17%	83.31%

Table 1: An overview of the gold standards datasets and their Inter-Annotator Agreement

We computed the IAA over the Ancient Greek-English and Ancient Greek-Portuguese datasets. Alignment agreement is considered in two cases, when both annotators align the same pair of tokens and when both annotators do not align a token. We also considered multi-word alignments (1-N, N-1, and N-N) as 1-1 pairs. For example, if the phrase ”The son” is aligned to ”υἱός”, it is converted to two 1-1 alignments (The, υἱός) and (son, υἱός). Let A_1 and A_2 be the flattened translation pairs created by the first and second annotators, respectively, and I is the intersection between them, we calculate the IAA as follows:

$$IAA = 2 * I/(A_1 + A_2)$$

Table 1 summarizes the IAA results and provides an overview of the gold standards datasets.

4. Automatic Word Alignment Model

We employ a state-of-the-art automatic word alignment workflow that utilizes pre-trained contextualized language models to generate word alignments. Further, we fine-tune a language model that can align Ancient Greek and English with a training strategy that combines unsupervised training over monolingual and bilingual datasets with supervised training over accurate alignments provided by UGARIT. The trained model significantly outperformed all statistical models such as Giza++, Elfomal, and fast_align on Ancient Greek-English and Ancient Greek-Portuguese parallel texts even with the absence of any training data for Ancient Greek-Portuguese.

4.1. Algorithm

Word alignment is the process of finding word-level equivalents between the source sentence $S = (s_1, s_2, .., s_n)$ and its translation $T = (t_1, t_2, .., t_m)$ (Brown et al., 1993). The alignment process takes S and T as inputs, and generate as output a set $A = \{(s_i, t_j) : s_i \in S, t_j \in T\}$ where s_i is a translation equivalent of t_j.

The core concept of recent studies (Jalili Sabet et al., 2020; Stengel-Eskin et al., 2019; Dou and Neubig, 2021) is to exploit the pre-trained multilingual contextualized language models such as mBERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2019) or fine-tuned versions of them. Then a similarity matrix can be derived based on distance/similarity metrics that calculate the similarity of the word embeddings for every two tokens. Then, the word-level alignments can be predicted by employing an extraction algorithm over this similarity matrix.

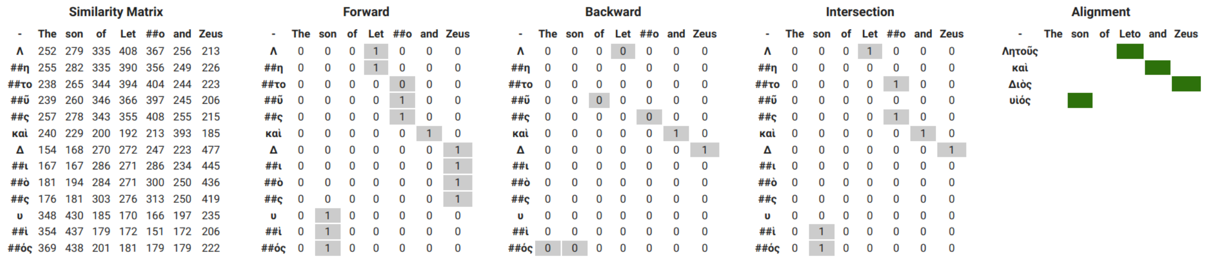
4.1.1. Similarity Matrix

Suppose S_{grc}, S_{eng} be two parallel sentences with lengths $n, m.$ and $SIM_{n \times m}$ the similarity matrix of these two sentences. Using the embeddings derived from multilingual transformers models, the similarity matrix can be filled as an equation 1:

$$\sum_{i}^{n}\sum_{j}^{m} SIM(i,j) = F_{sim}(t^{i}_{grc}, t^{j}_{eng}) \qquad (1)$$

Where t^i_{grc} is the embedding vector of the ith token in S_{grc}, t^j_{eng} is the embedding vector of the jth token in S_{eng}, and F_{sim} is a similarity function between the two vectors such as *Cosine Similarity*, *Dot Product*, or *Euclidean distance*. However, Our experiments showed that *Dot Product* slightly outperforms *Cosine Similarity*, and the results reported in tables 3 and 4 are based on *Dot Product* as a similarity function.

Regarding tokenization, mBert and XLM-R use different tokenizers; mBert uses a *WordPiece* tokenizer, whereas XLM-R uses a *byte-level BPE* tokenizer. Both tokenizers split words either into their full forms or into subwords (Figure 1). Moreover, we tested two embeddings alternatives for similarity matrices generation, the *subwords embeddings* provided by the language models and the *word-level embeddings* where each word is represented by the average embeddings of



Similarity Matrix

-	The	son	of	Let	##o	and	Zeus
Λ	252	279	335	408	367	256	213
##η	255	282	335	390	356	249	226
##το	238	265	344	394	404	244	223
##ῦ	239	260	346	366	397	245	206
##ς	257	278	343	355	408	255	215
καὶ	240	229	200	192	213	393	185
Δ	154	168	270	272	247	223	477
##ι	167	167	286	271	286	234	445
##ò	181	194	284	271	300	250	436
##ς	176	181	303	276	313	250	419
υ	348	430	185	170	166	197	235
##ὶ	354	437	179	172	151	172	206
##ός	369	438	201	181	179	179	222

Forward

-	The	son	of	Let	##o	and	Zeus
Λ	0	0	0	1	0	0	0
##η	0	0	0	1	0	0	0
##το	0	0	0	0	0	0	0
##ῦ	0	0	0	0	1	0	0
##ς	0	0	0	0	1	0	0
καὶ	0	0	0	0	0	1	0
Δ	0	0	0	0	0	0	1
##ι	0	0	0	0	0	0	1
##ò	0	0	0	0	0	0	1
##ς	0	0	0	0	0	0	1
υ	0	1	0	0	0	0	0
##ὶ	0	1	0	0	0	0	0
##ός	0	1	0	0	0	0	0

Backward

-	The	son	of	Let	##o	and	Zeus
Λ	0	0	0	0	0	0	0
##η	0	0	0	0	0	0	0
##το	0	0	0	0	0	0	0
##ῦ	0	0	0	0	0	0	0
##ς	0	0	0	0	0	0	0
καὶ	0	0	0	0	0	1	0
Δ	0	0	0	0	0	0	1
##ι	0	0	0	0	0	0	0
##ò	0	0	0	0	0	0	0
##ς	0	0	0	0	0	0	0
υ	0	0	0	0	0	0	0
##ὶ	0	0	0	0	0	0	0
##ός	0	0	0	0	0	0	0

Intersection

-	The	son	of	Let	##o	and	Zeus
Λ	0	0	0	1	0	0	0
##η	0	0	0	0	0	0	0
##το	0	0	0	0	1	0	0
##ῦ	0	0	0	0	0	0	0
##ς	0	0	0	0	1	0	0
καὶ	0	0	0	0	0	1	0
Δ	0	0	0	0	0	0	1
##ι	0	0	0	0	0	0	0
##ò	0	0	0	0	0	0	0
##ς	0	0	0	0	0	0	0
υ	0	0	0	0	0	0	0
##ὶ	0	1	0	0	0	0	0
##ός	0	1	0	0	0	0	0

Alignment

-	The	son	of	Leto	and	Zeus
Λητοῦς				🟩		
καὶ					🟩	
Διὸς						🟩
υἱός		🟩				

Figure 1: An overview of the alignment process, the similarity matrix is created with dot product function using the contextualized word embeddings of mBERT model.

its subwords. Table 5 shows that subwords embeddings always outperform word-level embeddings. In all experiments, the embeddings are extracted from the 8-th layer of mBERT and XLM-R since they have achieved the best performance.

Figure 1 shows an example of a similarity matrix computed using the *dot product* over the (sub)word embeddings extracted from the 8-th layer of mBERT.

4.1.2. Alignments Extraction

Once the similarity matrix is computed, alignments can be extracted by applying an extraction algorithm. (Dou and Neubig, 2021) proposed two probability thresholding-based methods to extract alignments from the similarity matrix, namely, $Softmax$ and $Entmax$ (Peters et al., 2019). Dou and Neubig (2021) apply the extraction in two directions and then considers the intersection between them (Figure 1).

Further, (Jalili Sabet et al., 2020) proposed three methods including $Argmax$, a baseline method, $Itermax$, an iterative method, and $Match$, a graph-based method. Jalili Sabet et al. (2020) found that $Itermax$ performs slightly better than $Argmax$, and it works better for distant languages, which was perfect for our case, since Ancient Greek and English are distant languages.

We employed the five extraction methods with their default settings in all our experiments. Section 4.4 compares and discusses in detail the performance of the various extraction methods. Figure 1 shows that the alignment is computed on subword-level. Since the task is to perform word-level alignment, we employed the heuristic principle *"two words are aligned if any of their subwords are aligned"* to convert subword-level alignments to word-level alignments similar to Jalili Sabet et al. (2020) and Dou and Neubig (2021).

4.2. Training Process and Objectives

The experiments we conducted on the pre-trained mBERT and XLM-R (Zero-Shot) showed significantly poor performance on both Ancient Greek-English and Ancient Greek-Portuguese datasets. Therefore, it was necessary to train and fine-tune those models aiming for better performance. Due to the availability of parallel sentences and in order to obtain the best outcome from the training process, we conducted several experiments employing multiple training objectives: *Masked Language Modeling (MLM)* (Gururangan et al., 2020), *Translation Language Modeling (TLM)* (Conneau and Lample, 2019), *Self-training Objective (SO)*, and *Parallel Sentence Identification (PSI)* (Dou and Neubig, 2021)[3]. Table 2 provides an overview of the conducted experiments and the employed training objectives and training datasets.

Experiment 1

In this experiment, we performed unsupervised fine-tuning of the mBERT and XLM-R using 32500 Ancient Greek-English parallel sentences with the training objectives MLM, TLM, SO, PSI. The parallel sentences used in this experiment are taken from Perseus Digital Library[4] (Iliad, Odyssey, Xenophon, New Testament).

Experiment 2

In this experiment, we also performed unsupervised training of the fine-tuned models of *Experiment 1*. We used 8000 Ancient Greek-Latin parallel Fragments with the same training objectives in the previous experiment. The parallel fragments are taken from the Digital Fragmenta Historicorum Graecorum project [5] (Berti, 2017).

Experiment 3

Multilingual contextualized language models mBERT and XLM-R are not trained on ancient Greek texts but on modern Greek, which is very different. So it was necessary to fine-tune them with monolingual Ancient Greek texts. In this experiment, we trained the fine-tuned models we obtained in *Experiment 2* on 12 million Ancient Greek tokens with Masked Language Model (MLM) training objective. The training dataset is extracted from Perseus Digital Library, the first thousand years of Greek project[6], and the PROIEL, PERSEUS[7], and Gorman[8] treebanking datasets.

[3] A detailed description of the training objectives is available in (Dou and Neubig, 2021)

[4] https://github.com/PerseusDL/canonical-greekLit

[5] https://www.dfhg-project.org/

[6] https://opengreekandlatin.github.io/First1KGreek/

[7] universaldependencies.org

[8] https://vgorman1.github.io/

Experiment	Input Models	Epochs	Training Objectives	Languages	Data Size	Source
EX 1	mBERT, XLM-R	1	MLM, SO, TLM, PSI	GRC-ENG	32.500 parallel sentences	Perseus
EX 2	EX1 fine tuned models	1	MLM, SO TLM, PSI	GRC-LAT	8.000 parallel sentences	DFHG
EX 3	EX2 fine tuned models	5	MLM	GRC Monolingual	12 Millions Tokens	Perseus, First1kGreek, TreeBanking
EX 4	EX2 fine tuned models	5	MLM, SO TLM, PSI	GRC-ENG GRC-LAT GRC-KAT	45.000 parallel sentences	Perseus, DFHG, UGARIT
EX 5	EX3 fine tuned models	5	MLM, SO TLM, PSI			
EX 6	EX5 fine tuned models	15	SO	Mixed dataset	2200 parallel sentences, 100k Translation Pairs	UGARIT

Table 2: An overview of the conducted experiments.

Experiment 4

This experiment and the next one aim to inspect the impact of training the models on monolingual texts; therefore, the two experiments use the same training data, but they differ by the pre-trained model. This experiment trains the model obtained in *Experiment 2* on 45000 parallel texts (We combined the datasets in *Experiments 1 and 2* with 4000 further parallel sentences taken from UGARIT database. The texts are in different languages mainly (Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Georgian).

Experiment 5

In this experiment, we train the model obtained after *Experiments 3* which is trained on monolingual Ancient Greek texts. For training, we use the same training dataset used in *Experiments 4* with the training objectives MLM, TLM, SO, PSI.

Experiment 6

In this experiment, we perform supervised training for the fine-tuned model obtained after *Experiment 5* using a word-level manually aligned dataset provided by UGARIT. The alignments are accurate and clean since they are done by scholars, teachers, and Experts. The dataset consists of 2265 parallel texts and almost 100k translation pairs.

4.3. Evaluation

4.3.1. Baseline Models

We compare our model to three popular statistical word alignment models, namely, Giza++, fast_align, elfomal. All these models require training data in the form of parallel sentences. We trained them on the 35000 Ancient Greek-English parallel sentences and 274 gold standard sentences and evaluated their performance on our produced gold standards. Tables 3 and 4 present the evaluation results. The poor performance on the Ancient Greek Portuguese dataset is because of the absence of the training data; the models are trained only on 183 sentences of the gold standard dataset.

4.3.2. Evaluation Metrics

Similar to (Och and Ney, 2003), we evaluate the performance of the alignment model against the gold standards, by employing $Precision$, $Recall$, $F1$, and Alignment Error Rate (AER) , which can be computed as in equations 2.

$$Precision = \frac{|A \cap P|}{|A|} \quad , \quad Recall = \frac{|A \cap S|}{|S|}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2}$$

$$AER = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|}$$

Where A indicates the alignments set predicted by the model, P and S indicate respectively the $Possible$ and $Sure$ alignment sets in the gold standards, and $|.|$ denotes the length of the set.

4.4. Results

The results reported in tables 3 and 4 are based on subwords embeddings and the *dot product* as similarity measure. The results show the superiority of the fine-tuned models over the statistical models (Giza++, elfomal, fast_align). Furthermore, *Experiments 5 and 6* show that the alignments derived from fine-tuned XLM-R models are superior to those derived from mBERT fine-tuned models for Ancient Greek/English and Ancient Greek/Portuguese datasets.

The results of *Experiment 2* indicate that the training the model on Ancient Greek-Latin parallel texts enhanced the alignment performance of Greek-English and decreased the AER by 0.79%-0.9% for XLM-R fine-tuned model and by 2.06%-2.42% for the fine-tuned mBERT model.

Moreover, The results indicate that training the model on monolingual data in *Experiments 3* decreased the $F1$ and increased the AER on both models. As mentioned before, in *Experiments 4 and 5*, we used the same training datasets and the same number of epochs. However, as table 3 shows, there is a significant difference in the results, as all the evaluation metrics values

		Precision		Recall		F1		AER	
Baseline	Giza++	37.25%		29.26%		32.78%		67.01%	
	FastAlign	37.37%		35.64%		36.48%		63.47%	
	Eflomal	47.17%		42.93%		44.95%		54.95%	
		mBERT				XLM-R			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
Zero-Shot	Softmax	37.14%	21.09%	26.90%	72.70%	37.59%	11.84%	18.01%	81.80%
	Entmax15	42.58%	17.34%	24.64%	74.94%	46.74%	8.67%	14.63%	85.20%
Ex1	Softmax	52.98%	38.21%	44.40%	55.28%	54.61%	28.21%	37.20%	62.46%
	Entmax15	56.96%	35.50%	43.74%	55.84%	65.35%	22.76%	33.76%	65.86%
Ex2	Softmax	55.89%	40.24%	46.79%	52.86%	55.62%	28.97%	38.10%	61.56%
	Entmax15	59.84%	37.04%	45.76%	53.78%	65.14%	23.49%	34.53%	65.07%
Ex3	Softmax	54.03%	39.68%	45.76%	53.94%	53.58%	24.21%	33.35%	66.33%
	Entmax15	60.35%	35.29%	44.54%	54.99%	61.99%	18.54%	28.54%	71.11%
Ex4	Softmax	65.06%	48.08%	55.30%	44.33%	65.22%	36.39%	46.72%	52.88%
	Entmax15	68.23%	46.01%	54.96%	44.58%	73.42%	30.34%	42.94%	56.57%
Ex5	Softmax	74.71%	54.61%	63.10%	36.50%	77.17%	54.42%	63.83%	35.81%
	Entmax15	77.60%	52.22%	62.43%	37.11%	82.84%	48.84%	61.45%	38.11%
	Match	58.09%	65.03%	61.36%	38.81%	62.46%	72.70%	67.19%	33.05%
	Argmax	78.72%	51.39%	62.18%	37.33%	84.73%	46.65%	60.17%	39.36%
	Itermax	69.64%	58.35%	63.50%	36.25%	77.07%	58.28%	66.37%	33.32%
Ex6	Softmax	80.80%	56.91%	66.78%	32.72%	90.73%	67.91%	77.68%	21.89%
	Entmax15	83.86%	53.76%	65.52%	33.93%	92.61%	64.18%	75.82%	23.69%
	Match	65.42%	**72.76%**	68.90%	31.31%	77.85%	**85.50%**	**81.50%**	**18.72%**
	Argmax	**84.95%**	52.47%	64.87%	34.57%	**93.44%**	62.57%	74.95%	24.54%
	Itermax	78.43%	64.08%	**70.53%**	**29.14%**	89.66%	72.05%	79.90%	19.73%

Table 3: Evaluation results on Ancient Greek-English gold standards.

		Precision		Recall		F1		AER	
Baseline	Giza++	25.59%		24.60%		0.2509		74.88%	
	fast align	25.62%		30.14%		27.70%		72.47%	
	Eflomal	34.84%		35.59%		35.21%		64.81%	
		mBERT				XLM-R			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
Zero-Shot	Softmax	30.08%	26.66%	28.27%	71.66%	21.68%	13.53%	16.66%	83.16%
	Entmax15	33.65%	22.67%	27.09%	72.73%	23.55%	9.27%	13.30%	86.44%
Ex6	Softmax	63.84%	61.27%	62.53%	37.40%	76.11%	75.61%	75.86%	24.13%
	Entmax15	65.49%	57.41%	61.18%	38.61%	**77.45%**	72.69%	74.99%	24.89%
	Match	50.00%	72.61%	59.22%	41.50%	58.79%	**86.17%**	69.89%	31.01%
	Argmax	66.01%	54.92%	59.96%	39.76%	77.25%	71.10%	74.05%	25.81%
	Itermax	59.67%	64.06%	61.79%	38.35%	72.22%	81.02%	**76.37%**	**23.91%**

Table 4: Evaluation results on Ancient Greek-Portuguese gold standards.

of *Experiment 5* are superior to *Experiment 4* values in both models with a large margin, which demonstrates the importance of training the models on monolingual datasets.

Table 4 shows that the proposed training strategy also achieved good results on Ancient Greek-Portuguese dataset with no supervised or unsupervised training on Ancient Greek-Portuguese parallel texts.

Regarding the alignment extraction approaches, the results show that $Itermax$ achieved the lowest AER and highest $F1$ in the majority of the experiments. $MWMF$ slightly outperformed $Itermax$ in *Experiments 5 and 6* based on fine-tuned XLM-R model in terms on AER. However, their Precision and Recall differ radically. As $MWMF$ achieved high Recall and low precision, $Itermax$ achieved opposite results. Further, $Softmax$ achieved better AER, $F1$, $Recall$ than $Entmax15$ in all experiments, while $Entmax15$ is always superior regarding $Precision$.

To inspect the output of the different alignment extraction approaches, we used the visual evaluation tool proposed by (Yousef and Jänicke, 2022). Figure 2 shows that *MWMF* generates the highest number of translation pairs among the other extraction approaches, which explains why *MWMF* consistently achieves the highest Recall and the lowest Precision as reported in tables. Moreover, *Argmax* generates the lowest number of translation pairs. Therefore, it achieves always the lowest Recall and the highest Precision.

4.4.1. Qualitative Evaluation

In addition to the quantitative evaluation, we conducted a qualitative evaluation on 50 Greek-English sentences and 52 ancient Greek-Portuguese sentences for detailed observation of the errors and mismatch patterns.

For Greek-English alignments, from 527 translation pairs, there were 30 incorrect pairs (5.6%). The most frequent inaccuracy was in partially correct pairs. They

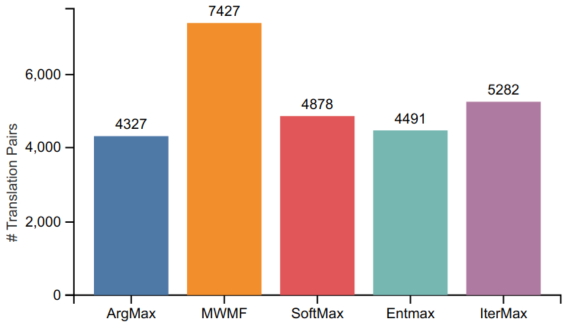


Figure 2: A comparison among the alignment extraction approaches regarding the number of translation pairs they produce based on the XLM-R fine-tuned model obtained in Experiment 6.

included an unrelated token that should have been excluded, or they were missing a token in the English translation to match correctly. This partially incorrect matching counts for 20 instances out of 30 errors. Of this number, 13 instances included an extra token in the English translation (Example: ἄνακτος: lord strikes), two instances included an extra token in the Greek (example: ἐξεπράθομεν δέδασται : we apportioned), and five instances were missing token(s) in the English translation (Example: ἐΐκτην : were, missing the token "like" in the English translation). The rest of the errors (10 out of 30) were tokens that did not match at all, most commonly, particles and conjunctions (5 instances) that are often difficult to translate or align. The rest of the errors can be attributed to the complexity and ambiguity of the translation or the text.

For Greek-Portuguese alignments, from 509 translation pairs, there were 51 incorrect pairs (10%). As in Greek-English, the most frequent inaccuracy (24 out of 51 errors) was found in pairs that were partially correct (e.g.: ἄλλοι: os outros; ὄσσε: olhos pareciam; οὕτω: tão cedo). Another 11 mismatches were related to missing a word in Portuguese translation (e.g.: γλυκίων: mais; καρτερός: vigoroso). Among the 14 tokens that were not a complete match, the inverted alignments seemed to occur where they were in close proximity either in the original or in the translation (e.g.: φηρσὶν: os monteses, and ὀρεσκῴοισι: centauros). Also, the ambiguity of the Portuguese term ”a” (preposition and definite article) may have caused another type of error (e.g. πρὸ: a). In general, the errors appeared to be more critical when Portuguese shows a two-word comparative, periphrastic, or multi-word expression translation.

4.5. Conclusion

Tables 3 and 4 show that fine-tuning the multilingual models on monolingual datasets played a key role in enhancing the performance of the model. The results can be further enhanced by performing supervised and unsupervised fine-tuning even on language pairs that are different from the target language pairs such as Ancient Greek-Italian and Ancient Greek-French.

Moreover, if monolingual or bilingual texts are available, we recommend fine-tuning the model on monolingual data with Masked Language Modeling training objective first, then performing supervised fine-tuning with the desired training objectives.

The results also show the great impact of supervised training with word-level manual alignments which decreased the AER from 31.85% to 33.05% on fine-tuned XML-R model and from 36.25% to 29.14% on fine-tuned mBERT model. This leads us to conclude that fine-tuned XMl-R models are more sensitive to supervised fine-tuning than fine-tuned mBERT models.

5. Future Work

Our research showed that fine-tuning pre-trained multilingual contextualized models on both mono- and bilingual training datasets significantly impact the accuracy of automatic alignment models. However, we plan to enhance the model further:

- *Unsupervised Mono- and Multi-lingual Training:* The current model is trained on 12 million Ancient Greek tokens and 45000 bilingual sentences. We intend to expand the corpus by thoroughly inspecting available digital libraries to collect more texts in Ancient Greek that are reliably aligned at paragraph or sentence level with their translations.
- *Supervised Multi-lingual Training:* We were keen to train a high-quality alignment model; therefore, the current model is trained on only ten trusted UGARIT users' manual alignments. We dismissed all other alignments created by students or non-identified users. We will evaluate the existing alignments of Ancient Greek, select the accurate ones, and include them in the training data.
- *Alignment Extraction Approach:* We are developing an alignment extraction algorithm that combines the high Recall of $MWMF$ and the high Precision of $Itermax$.

Further, we intend to create alignment gold standards for other language pairs such as Ancient Greek/Latin, Ancient Greek/Italian, and Ancient Greek/Persian and test the model performance on the new datasets. Finally, The great performance achieved by fine-tuning the model on word alignment task has encouraged us to test the model on other downstream tasks such as named-entity recognition and Part-Of-Speech tagging of Ancient Greek texts.

6. Acknowledgements

We are grateful to the UGARIT community of experts and annotators, who provided the necessary data and support for this research. Moreover, we thank scholars André Malta and Emerson Cerdas for providing the Portuguese translations of the texts used in our corpus; and Gregory Crane, for the guidance and support of the UGARIT project.

7. Bibliographical References

Baker, M. (2000). Towards a Methodology for Investigating the Style of a Literary Translator.

Bamman, D., Mambrini, F., and Crane, G. (2010). An ownership model of annotation: The ancient greek dependency treebank. 12.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Burnet, J. (1903). *Platonis opera. T. 3: Tetralogias V - VII continens*. Scriptorum classicorum bibliotheca Oxoniensis. Clarendon, Oxonii, nachdr. edition.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Dagan, I., Church, K., and Gale, W. (1999). Robust Bilingual Word Alignment for Machine Aided Translation. In Susan Armstrong, et al., editors, *Natural Language Processing Using Very Large Corpora*, Text, Speech and Language Technology, pages 209–224. Springer Netherlands, Dordrecht.

David, Y., Grace, N., Richard, W., et al. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8.

DeNero, J. and Klein, D. (2007). Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dou, Z.-Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online, April. Association for Computational Linguistics.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July. Association for Computational Linguistics.

Jalili Sabet, M., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November. Association for Computational Linguistics.

Kay, M. and Röscheisen, M. (1993). Text-translation Alignment. *Comput. Linguist.*, 19(1):121–142, March.

Marchant, E. C. (1910). *Xenophontis Opera omnia*. Clarendon Press. OCLC: 802674413.

Melamed, I. D. (1998). Manual annotation of translational equivalence: The blinker project. *arXiv preprint cmp-lg/9805005*.

Müller, M. (2017). Treatment of markup in statistical machine translation. Association of Computational Linguistics.

Murray, A. (1924). *The Iliad. With an English translation by A.T. Murray*. William Heinemann ; G.P. Putnam's Sons London (England) : New York (New York).

Nicolai, G. and Yarowsky, D. (2019). Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1765–1774, Florence, Italy, July. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, page 440–447, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Padó, S. and Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Palladino, C., Foradi, M., and Yousef, T. (2021). Translation alignment for historical language learning: a case study. *Digital Humanities Quarterly*, 15(3).

Peters, B., Niculae, V., and Martins, A. F. (2019). Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519.

Steingrímsson, S., Loftsson, H., and Way, A. (2021). CombAlign: a tool for obtaining high-quality word alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 64–73, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Stengel-Eskin, E., Su, T.-r., Post, M., and Van Durme, B. (2019). A discriminative neural model for cross-lingual word alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 910–920, Hong Kong, China, November. Association for Computational Linguistics.

Jean Véronis, editor. (2000). *Parallel Text Processing: Alignment and Use of Translation Corpora*. Text, Speech and Language Technology. Springer Netherlands, Dordrecht-Boston-London.

Werner, C. (2018). *Homero, Ilíada*. Ubu Editora, 1ª edição edition, November.

Yousef, Tariq, P. C. S. F. and Foradi, M. (forthcoming 2022). Translation alignment with ugarit. *Information*.

Yousef, T. and Jänicke, S. (2022). Visual evaluation of translation alignment data. In *Proc. EuroVis*, volume 22.

Yousef, T., Palladino, C., Shamsian, F., and Foradi, M. (2022). Translation alignment with ugarit. *Information*, 13(2).

8. Language Resource References

Berti, M. (2017). Digital fragmenta historicorum graecorum (dfhg).

Graça, J., Pardal, J. P., Coheur, L., and Caseiro, D. (2008). Building a golden collection of parallel multi-language word alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Holmqvist, M. and Ahrenberg, L. (2011). A gold standard for English-Swedish word alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 106–113, Riga, Latvia, May. Northern European Association for Language Technology (NEALT).

Kruijff-Korbayová, I., Chvátalová, K., and Postolache, O. (2006). Annotation guidelines for czech-english word alignment. In *LREC*, pages 1256–1261. Citeseer.

Lambert, P., DE GISPERT, A., BANCHS, R., and MARINO, J. B. (2005). Guidelines for word alignment evaluation and manual alignment. *Language resources and evaluation*, 39(4):267–285.

Macken, L. (2010). An annotation scheme and gold standard for dutch-english word alignment. In *7th conference on International Language Resources and Evaluation (LREC 2010)*, pages 3369–3374. European Language Resources Association (ELRA).

Mareček, D. (2008). Automatic alignment of tectogrammatical trees from czech-english parallel corpus. Master's thesis, Charles University, MFF UK.

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, pages 1–10.

Appendix

	mBert		XLM-R	
Tokenization	Subword	Word	Subword	Word
Softmax	32.72%	34.47%	21.89%	24.07%
Entmax	33.93%	35,81%	23,69%	25,42%
Match	31.31%	37.32%	18.72%	29.06%
Argmax	34.57%	36.29%	24.54%	26.16%
Itermax	29.14%	29.62%	19.73%	19.14%

Table 5: A comparison between two different embeddings alternatives: the subword embeddings as they are provided by the language models, and the word-level embeddings which is calculated for wach word as the average embeddings of its subwords. This comparison is based on the embeddings provided by the Greek-English fine-tuned models obtained in Experiment 6.

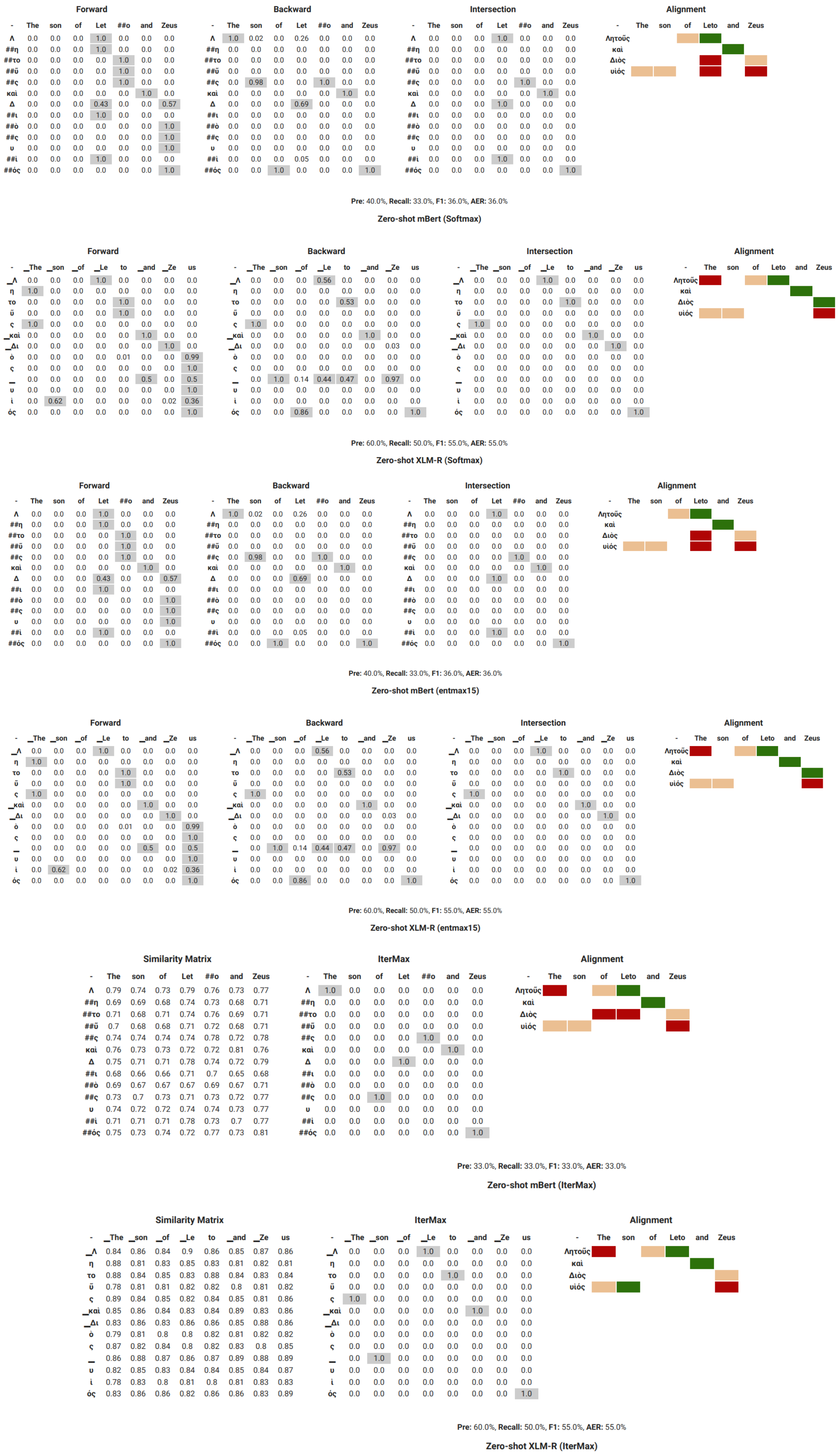


Figure 3: A Comparison among different alignment extraction approaches from similarity matrices derived from pre-trained mBERT and XLM-R. With the evaluation results on Gold Standards.
Color Codes: Green (Correct alignment), Red (Incorrect Alignment), Orange (Missing Alignment).

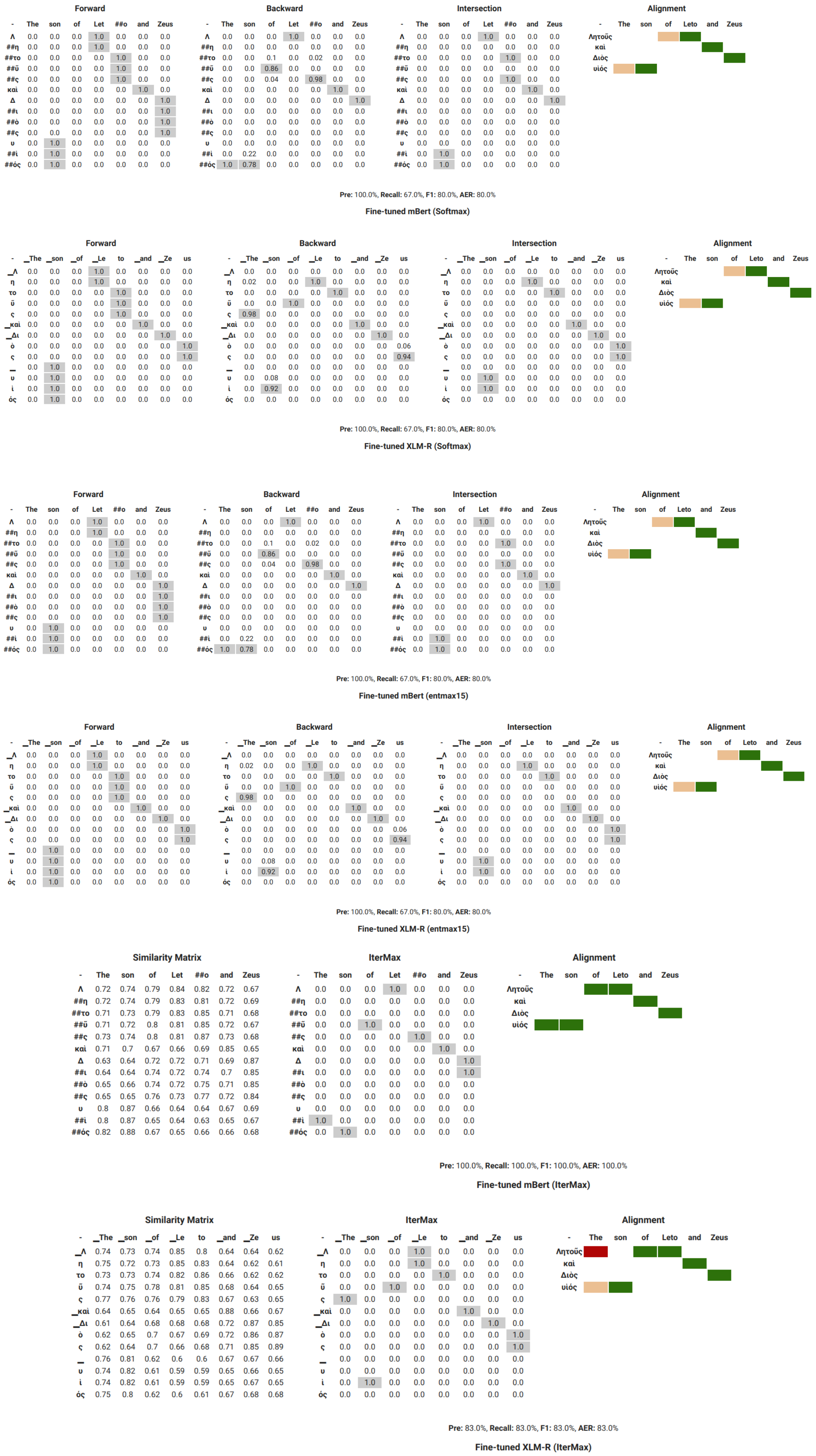

Forward
Backward
Intersection
Alignment
Pre: 100.0%, Recall: 67.0%, F1: 80.0%, AER: 80.0%
Fine-tuned mBert (Softmax)
Forward
Backward
Intersection
Alignment
Pre: 100.0%, Recall: 67.0%, F1: 80.0%, AER: 80.0%
Fine-tuned XLM-R (Softmax)
Forward
Backward
Intersection
Alignment
Pre: 100.0%, Recall: 67.0%, F1: 80.0%, AER: 80.0%
Fine-tuned mBert (entmax15)
Forward
Backward
Intersection
Alignment
Pre: 100.0%, Recall: 67.0%, F1: 80.0%, AER: 80.0%
Fine-tuned XLM-R (entmax15)
Similarity Matrix
IterMax
Alignment
Pre: 100.0%, Recall: 100.0%, F1: 100.0%, AER: 100.0%
Fine-tuned mBert (IterMax)
Similarity Matrix
IterMax
Alignment
Pre: 83.0%, Recall: 83.0%, F1: 83.0%, AER: 83.0%
Fine-tuned XLM-R (IterMax)

Figure 4: A Comparison among different extraction methods on the similarity matrices derived from fine-tuned models.